Disambiguating the Context of the Concept Terms using Concept Hierarchies

Raju Dara, T. Raghunadha Reddy

Abstract*: *Latent Semantic Analysis (LSA) makes the machine clearly conceptualize the terms of the document by learning the context in which these terms are written. However, LSA suffers from the limitation of input data matrix size in terms of number of terms and number of documents of the considered dataset. When the size of the dataset is huge, LSA becomes inefficient towards learning the correct context and thereby is unable to produce the intended concepts by the machine. To overcome this problem, Context Disambiguation (ConDis) ontology is engineered for a domain which has the capability of evolving itself based on automatic learning of concepts and relations from the ever scaling documents over the web. The concept hierarchies from general to specific concepts combined with corresponding object relations specify the particular context for a term. These object relations based concept hierarchies clearly help disambiguate the context of the concept terms in an effective manner.

Index Terms*: *LSA, term, context, ConDis ontology, concept, ontology evolution, concept hierarchies.

I. INTRODUCTION

In the perspective of document understanding, LSA is used to extract the semantics of the document on the basis of word associations [1]. LSA is a mathematical model which is free from outside sources like vocabularies, dictionaries and so on [2]. Humans articulate various thoughts using language. During the language evolution, scripts were developed for language representation. Then rules of language grammar were defined to administer the word groups towards creating the meaning among the words in a document. In effect, the meaning of a document relies on co-occurrences of the words written in a language.

The learning of semantic information begins from understanding the core words in the language. Consequently it is important to represent the relationships between words in the documents to realize the semantics. LSA is used for this purpose. It also ascertains the groupings between words that coexist within similar contexts. It is thus recommended as a form for understanding the language.

LSA makes the machine clearly conceptualize the terms of the document by learning the context in which these terms are written. The maximum number of rows and columns used for input matrix data size in LSA based document collection analysis was 10783 × 600 [3]. LSA was unable to process the documents beyond this matrix size. When the count of documents in the dataset grows exponentially, LSA becomes inefficient towards learning the correct context and thereby it is unable to produce the intended concepts by the machine.

To overcome this problem, a ConDis ontology is engineered for a domain which has the capability of evolving itself based on automatic learning of concepts and relations from the ever scaling documents over the web. The concept hierarchies from general to specific concepts combined with corresponding object relations specify the particular context for a term. These object relations based concept hierarchies clearly help disambiguate the concept terms in an effective manner.

II. RELATED WORKS

LSA was surveyed under various research areas after its proposal. Berry et al. used [4] for information retrieval. The researchers used Singular Value Decomposition (SVD) to assess the word usage structures among the documents. It was performed using vectors acquired from SVD. It was observed that statistically obtained vectors were more strong indicators for meaning of the document when compared with the terms individually.

Landauer et al. used [5] semantic spaces of LSA to model the associations elucidated in the memory of the humans. Based on word co-occurrences, the LSA space was used to discover the knowledge required to pass a word synonym test. Using this method LSA resulted in 64% accurate for the standard student taking test and 64% accuracy for word projection methods.

In the perspective of document classification, the LSA performance relies on the amount of arrangement of axes to the new determinate position. This reorientation shifts the training documents closer to the test document of the appropriate concept. Recently, Karthik Krishnamurthi et al. improved [6] the performance of the LSA in terms of concept term disambiguation by supplying supplementary and summary information of the considered data corpus to LSA.

Ontology development is a less tackled in the area of research. Heflin mentioned [7] the ontologies need in distributed and dynamic environments and provided the ontologies formal definition. Ontology versioning is analysed in the work of [8]. The researchers presented the consequences and causes of the fluctuations in the ontology development.

Revised Manuscript Received on October 05, 2019.

Dr. Raju Dara, Department of CSE, Vignana Bharathi Institute of Technology, Hyderabad, Telangana, India. Email: rajurdara@gmail.com

Dr. T Raghunadha Reddy, Dept of Information Technology, Vardhaman College of Engineering, Shamshabad, Telangana, India.
Email: raghu.sas@gmail.com

Maedche et al. introduced [9] the idea of ontology evolution based on the changes in requirements of users. In the work of [10] the researchers presented the principles for guiding, building reliable and correct ontologies.

In all the above works, both formal and informal context models were proposed. The informal context models depend on copyrighted representation schemes. Context Toolkit [11] and Cooltown [12] were the developed informal context models. Formal context models utilize the approaches of formal modeling to work on context. Karen et al. modeled [13] the context using both UML and ER models. Anand et al. represented [14] the context as first-order predicates in Gaia system.

As mentioned earlier, LSA makes the machine clearly conceptualize the terms of the document by learning the context in which these terms are written. However, LSA suffers from the limitation of input data matrix size in terms of terms count and documents count of the considered dataset. When the size of the dataset is huge, LSA becomes inefficient towards learning the correct context and thereby is unable to produce the intended concepts by the machine [15].

So, a separate ontology is required as the ontology has the capacity to scale itself based on the evolution that occurs over the time period. This scalable ontology never restricts the size on the input document collection. Finally, the existing formal context models support certain level of context reasoning. Also the previous works except [16] on ontology based contextual reasoning has never addressed formal knowledge sharing, or shown the feasibility of context reasoning. Semantic Web Rule Language (SWRL) rules are to be devised upon the ontology constructs to show the feasibility of context reasoning. These rules clearly express the context of the concept term.

III. ONTOLOGY BASED METHOD

The principal objective of disambiguating the concept terms from scalable documents by the machine is to utilize the concept hierarchies that are obtained by both engineering and evolving the ontology over time. For achieving this target, a framework is provided in Figure 1.

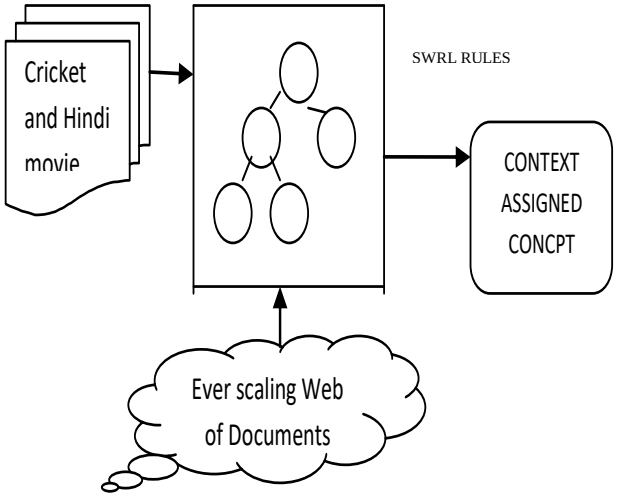

Fig. 1. Proposed Framework

The framework is organized of two modules. The first is the development of the Concept Disambiguation (ConDis) Ontology in terms of a specific ontology. The ConDis ontology is a domain ontology that defines the details of concepts of CONON upper ontology. Next, the engineered ConDis ontology is evolved based on the incoming documents from the web. This evolution of ontology is carried out to eliminate the input data processing issues inherent in LSA approach. The ontology evolution is based on automatic learning of concepts and relations from these documents. The second and final module is to assign the exact context to the concept terms based on the object relations available on the concept hierarchies.

3.1 Engineering the ConDis ontology

A. The domain and scope of ConDis ontology

The ConDis Ontology covers the domain of texts of English language. The ConDis Ontology is used to relate concepts and data properties to learn the various situations. ConDis Ontology answers the question on the concepts leading to learning the target context.

B. Classes in the ConDis ontology

In ConDis Ontology, the top down approach is used to define the classes and the created general classes are Person, Entertainment, Game, Movie, and so on. The specialised classes namely Outdoor, Indoor, Regional, National, Cricket and Hindi were created.

C. Classes arrangement in subclass–superclass hierarchy

In hierarchical manner the classes were organized in the ConDis ontology. The taxonomy of ConDis Ontology class is displayed in Figure 2.

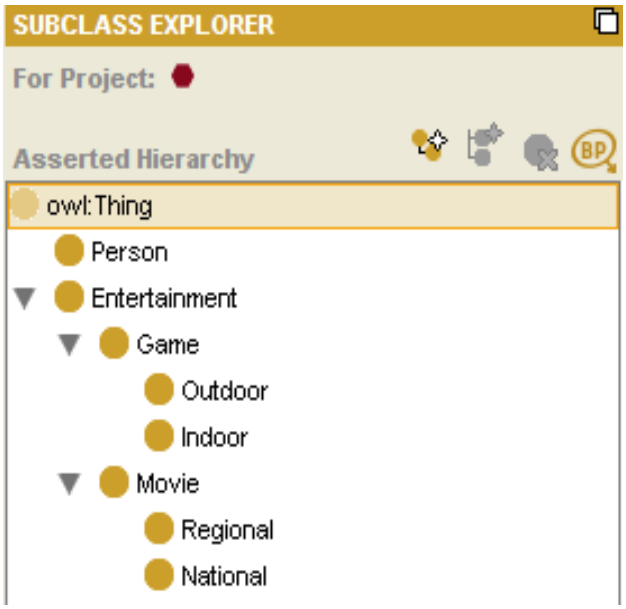

Fig. 2. ConDis ontology class taxonomy

D. Slots description

The classes defined in ConDis Ontology provide less data to answer the competent questions. In the time of defining of classes, the internal relationships among the classes also described. These relationships were called the properties of the class. The properties of object are namely belongsTo, speaksLanguage, plays, actsIn etc. The object properties are shown in Figure 3.

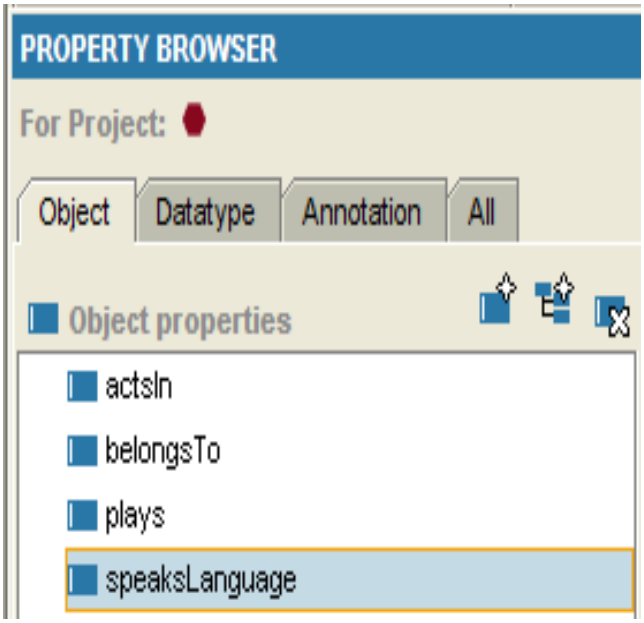

Fig. 3. ConDis object properties

E. The creation of slot instances

The last step is creating instances individually for the classes specified in the hierarchy. The class Person in Figure 4 is defined with the item instance and the related instances are specified from the range class of the relationship.

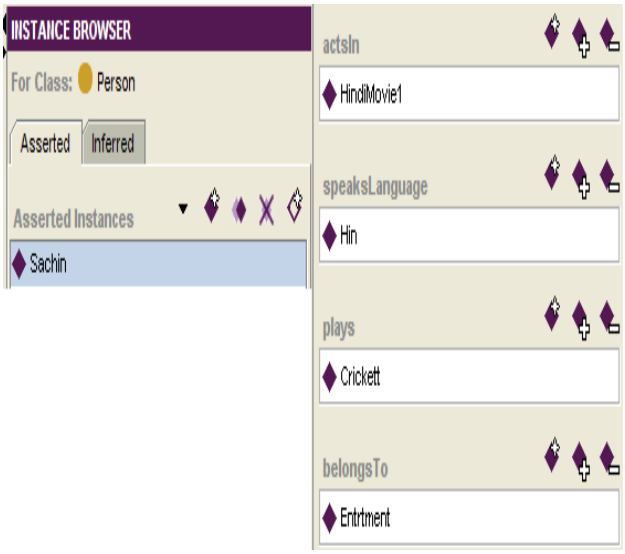

Fig. 4. Person instance and its slots values

The ConDis Ontology is visualized in the form of tree as given below.

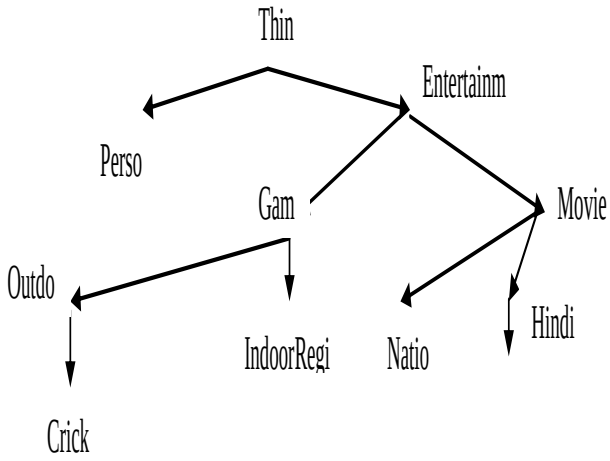

Fig. 5. Visualization of ConDis ontology

3.1.1 Ontology evolution based on automatic learning of concepts and relations from documents

Ontology Evolution is the revision of the ontology timely to meet the changed design and requirements of business. A modification in one part of the ontology must and should maintain the consistency of the entire ontology. A careful analysis of various types of ontology changes helps to ensure automatic updation of concepts and relations so that the evolution process is well realized.

Table- I: Ontology Evolution change methods

S.No.	Change	Method
1	Add	add_subconcept add_superconcept add_concept add_subconceptof add_domain add_axiom
2		
3		
4		
5		
6		
7	Delete	delete_concept delete_subconceptof delete_domain delete_axiom
8		
9		
10		
11	Merge	merge_concepts
12	Extract	extract_subconcepts
13		extract_superconcept
14		extract_relatedconcept

The automatic learning of concepts and relations is carried out in the sequential manner. First, the named entities from the documents are learned using the vocabulary of DBpedia, a linked data version of Wikipedia web document collection. Then, the concepts of the identified named entities are obtained using the corresponding Wikipedia document class.

The relations among the concepts are identified by semantic parsing the sentences and resolving for co-references first. Then, by using the predicates of DBpedia vocabulary, the corresponding relations are identified.

The identified concepts and relations are carefully added to the ontology by resolving the changes and keep the ontology consistent. The changes are represented in Table I.

Once all the required changes have been carried out, the ontology is verified for semantics of change, approval of these changes and propagating these changes.

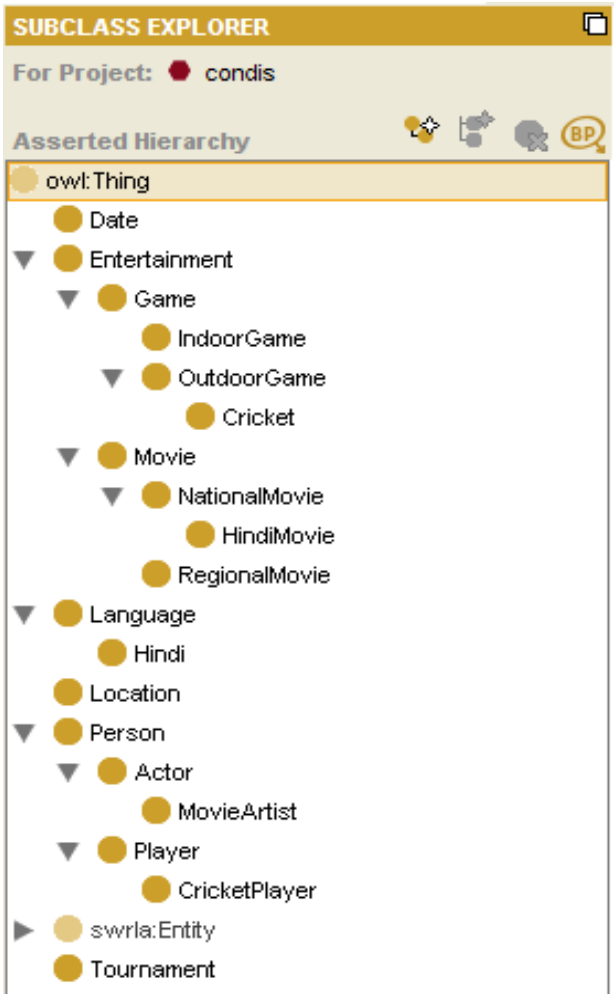

Fig. 6. Evolved ConDis ontology

The movie sub-ontology in the ConDis is evolved in the following way:

a)add_subconcept: The concepts namely Cinema, Genre and Trailer are added as sub concepts under Entertainment concept. These concepts stand at the level of Movie concept. The game sub-ontology in the ConDis is evolved in the following way:

b)add_concept: The concepts namely Date, Location, Player and Tournament are added as concepts under Thing concept.

c)add_subconceptof: Move the concept Player and make it as a child concept under Person using add_subconceptof.

Finally, the evolved ConDis ontology is illustrated in the figure 6.

3.2 Context assignment to concept terms with object relations and concept hierarchies from ConDis ontology.

When the context is modelled using the formal approach, the logical reasoning mechanisms is used to process the context. The context reasoning has two advantages such as the context consistency is checked and deducing implicit context (the actual context assignment) from explicit context (the concept hierarchies and object properties between the concepts).

A sports person and artist disambiguation scenario is presented to describe the context reasoning role in ontology based computing. By defining concept hierarchies and relating these concepts, the term is possible to be disambiguated. For example, when the person belongs to Entertainment and is part of the Game and plays the outdoor game and the game is cricket, then the person is a cricket player. Similarly, when the person belongs to Entertainment and is part of the movies and acts in national movies and the language is hindi, then the person is a movie artist. It is obvious that high-level context cannot be directly acquired from ontology. It is reasoned using concept driven low-level context such as relations among the concepts.

The object properties and the data properties devised as links between the concepts provide the necessary contextual information for disambiguating the concept. The concept hierarchies also help in the concept disambiguation process.

The SWRL rules are written by using the constructs of the ontology. These rules assign the context to the concept term by using the object relations and concept hierarchies. Following are the SWRL rules that provide the context to the concept term thereby disambiguating the term in the efficient manner.

belongsTo(?x, ?y) ∧ plays(?x, ?p) ∧ Cricket(?q) → CricketPlayer(?x)

belongsTo(?x, ?y) ∧ actsIn(?x, ?p) ∧ speaksLanguage(?x, ?q) → MovieArtist(?x)

belongsTo(?x, ?y) ∧ plays(?x, ?p) → Player(?x)

belongsTo(?x, ?y) ∧ actsIn(?x, ?p) → Actor(?x)

The debugged SWRL rules are shown in below figure.

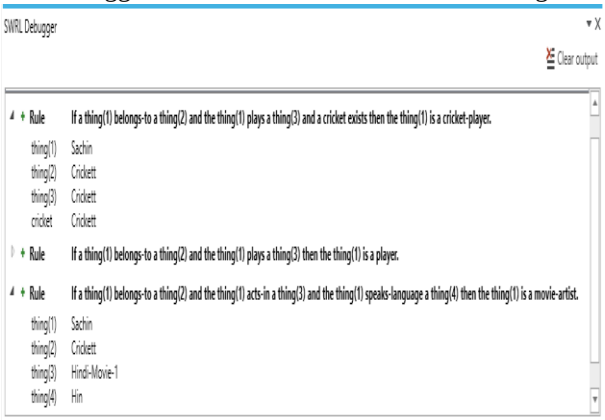

Fig. 7. Debugged SWRL rules

IV. RESULTS AND DISCUSSIONS

This section analyzes and presents the results of the conducted experiment with context reasoning. The aims of this experiment are to deliver the information which is helpful for implementing the context reasoning in game and movie environments.

The description logic based ontology reasoning is used to carry out this experiment. Ontology reasoning was generated by using Jena2 Semantic Web Toolkit [17], which helps to support rule based inference on OWL graphs.

To analyze the context on large-scale dataset, the ConDis and CONON Upper Ontology are combined to generate several context datasets. These datasets range from small-scale of 1K RDF triples to large-scale of above 10K RDF triples.

The dataset size is measured in terms of the count of RDF triples and each triple represents a single Subject – Predicate - Object (S-P-O). For the considered cricket documents and hindi movie documents, the dataset is contained with 1357 OWL classes and instances of 3264 RDF triples. The Current version of ConDis has 19 OWL classes. This is seen as a small-scale context dataset. The CONON Upper Ontology contains 2534 classes which is a large-scale context dataset.

The number of concept terms disambiguated using ConDis ontology are 786 out of 816 concept terms. This means, the precision of concepts assigned with correct context is 96%. The precision obtained using the current approach is much higher than the Roman et al.[18] work on ontology based word sense disambiguation. The researchers have achieved a precision of 88.82%. The performance of the ConDis Ontology in terms of classifier is compared with various parameters against the work of [18] in the table below.

This is because the researchers in their work [18] have restricted to analyzing the co-occurrences between the concepts to define the context of a concept. However, in the current work the concept hierarchies and the relations between the concepts were considered to disambiguate the context of the concept. Also the ontology used in [18] is not an evolutionary ontology. This cannot guarantee the disambiguation of all the concept terms as the knowledge embodied in their ontology is less.

Table- II: Ontology as classifier performance parameters

Classifier used	Accuracy	Evolutionary Ontology	Usage of Concept hierarchies
Ontology based WSD [17]	88.82%	No	No
Ontology based SWRL rules (our work)	96%	Yes	Yes

V. CONCLUSION

The disambiguation of the context of the concept term using ConDis ontology hierarchies and relations was carried out successfully. The ConDis ontology was an evolutionary one. This provides the machine to clearly disambiguate the terms in an efficient manner.

ACKNOWLEDGEMENT

I will be grateful to the FIST laboratories for providing required computational facilities at the institution, and I extend my deep sense of gratitude to the management, principal, director of R&D of Vignana Bharathi Institute of Technology for the accorded support.

REFERENCES

1. Deerwester S., Dumais S., Furnas G. and Landauer T. K., "Indexing by latent semantic analysis", American Society for Information Science, 1990, 391–407.
2. Landauer T. K. and Foltz P. W., "An Introduction to Latent Semantic Analysis", Discourse Processes, 1998, 259–284.
3. Krishnamurthi, K., Sudi, R. K., Panuganti, V. R., & Bulusu, V. V. (2013, August). An Empirical Evaluation of Dimensionality Reduction Using Latent Semantic Analysis on Hindi Text. In Asian Language Processing (IALP), 2013 International Conference on (pp. 21-24). IEEE.
4. M. Berry and S. Dumais, "Using linear algebra for intelligent information retrieval", SIAM Review, pp. 573–595, 1995.
5. Landauer T.K. and Dumais S.T., "Latent semantic analysis and the measurement of knowledge", Educational Testing Service Conference on Natural Language Processing Techniques and Technology in Assessment and Education, 1994.
6. Krishnamurthi, Karthik, Vijayapal Reddy Panuganti, and Vishnu Vardhan Bulusu. "Understanding Document Semantics from Summaries: A Case Study on Hindi Texts." ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP) 16.1 (2016): 7.
7. J. Heflin, Towards the Semantic Web: Knowledge Representation in a Dynamic, Distributed Environment, Ph.D. Thesis, University of Maryland, College Park. 2001.
8. M. Klein and D. Fensel, Ontology versioning for the Semantic Web, Proc. International Semantic Web Working Symposium (SWWS), USA, July 30 – August 1, 2001.
9. A. Maedche and S. Staab, Ontology Learning for the Semantic Web, IEEE Intelligent Systems, 16(2), March/April 2001. Special Issue on Semantic Web, 2001.
10. D. McGuinness, Conceptual Modeling for Distributed Ontology Environments, In the Proceedings of the ICCS 2000, August 14-18, Darmstadt, Germany , 2000.
11. A.K.Dey, et al. A Conceptual Framework and a Toolkit for Supporting the Rapid Prototyping of Context-Aware Applications- , Human-Computer Interaction Journal 16(2-4), pp. 97-166, 2001.
12. Tim Kindberg, et al. People, Places, Things: Web Presence for The Real World, Technical Report HPL-2000-16, HP Labs, 2000.
13. Karen Henricksen, et al.,Modeling Context Information in Pervasive Computing Systemsi , Pervasive 2002.
14. Anand Ranganathan, et al. A Middleware for Context-Aware Agents in Ubiquitous Computing Environmentse,USENIX International Middleware Conference, 2002.
15. Raghunadha Reddy T, Vishnu Vardhan B, Vijayapal Reddy P, "A Survey on Author Profiling Techniques", International Journal of Applied Engineering Research, March 2016, Volume-11, Issue-5, pp. 3092-3102.
16. Wang, X. H., Zhang, D. Q., Gu, T., & Pung, H. K. (2004, March). Ontology based context modeling and reasoning using OWL. In Pervasive Computing and Communications Workshops, 2004. Proceedings of the Second IEEE Annual Conference on (pp. 18-22). Ieee.
17. Jena2 Semantic Web Toolkit: http://www.hpl.hp.com/semweb/jena2.htm
18. Prokofyev, R., Demartini, G., Boyarsky, A., Ruchayskiy, O., & Cudré-Mauroux, P. (2013, March). Ontology-based word sense disambiguation for scientific literature. In European conference on information retrieval (pp. 594-605). Springer, Berlin, Heidelberg.

AUTHORS PROFILE

Dr. Raju Dara is a professor of Computer Science and Engineering Department at Vignana Bharathi Institute of Technology, Hyderabad. He has 16 years of teaching experience for Graduate and Post Graduate engineering courses. His current research interests are Data Warehousing, Image Processing, and Network Security. He published 30 research papers in international journals as well as international conferences.

Dr. T. Raghunadha Reddy, working as Associate Professor in the Department of Information Technology, Vardhaman College of Engineering, Shamshabad, Hyderabad, Telangana, India. He has 15 years of teaching experience. His current research interests are Data mining, Natural Language Processing, Information Retrieval and Text Classification. He published 40 research papers in reputed international journals and international conferences. He has memberships in IET and IAENG.